CERAC

Centro de Recursos para el Análisis de Conflictos

www.cerac.org.co

Documentos de CERAC

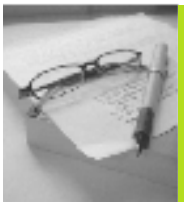
ISSN: 1909-1397

N° 13

On the Performance of Dual System Estimators of Population Size: A Simulation Study

Mauricio Sadinle

Diciembre, 2008



CERAC
Centro de Recursos para el Análisis de Conflictos

El Centro de Recursos para el Análisis de Conflictos (CERAC) es un centro de investigación privado e independiente, especializado en la generación de recursos para la investigación sobre violencia, el análisis de conflictos armados y el estudio de sus impactos sobre el desarrollo socioeconómico y el bienestar de las personas. CERAC no tiene filiación partidista o gubernamental alguna. CERAC busca contribuir a la reducción de la violencia y su impacto sobre las personas, las comunidades y la sociedad, mediante la investigación en ciencias sociales basada en evidencia, el desarrollo de tecnologías e innovación, además de participar en el diseño, implementación y evaluación de políticas públicas e intervenciones dirigidas a reducir la violencia armada. (Para mayor información visítese: http://www.cerac.org.co/acercade.htm)

The Conflict Analysis Resource Center (CERAC) is a private and independent think tank, focused in the generation of resources to study violence, armed conflicts and their impact on social and economic development and people's welfare. CERAC does not have any partisan or governmental affiliation. The Center aims to contribute to the reduction of violence and its impact on individuals and communities, through social science research based on verifiable information; the development of technologies and innovation, and the involvement in the design, implementation and evaluation of public policies and interventions. (For more information, visit: www.cerac.org.co/aboutus.htm)

La Serie Documentos de CERAC son publicaciones ocasionales de los resultados preliminares y avances de investigación de los miembros de CERAC, sus investigadores asociados, o de investigadores que han hecho uso de los recursos del Centro. Estos documentos son trabajo en curso, y como tal, están sujetos a modificaciones. Sus autores agradecen la retroalimentación y comentarios de los lectores.

The Working Papers Series of CERAC are occasional publications of preliminary research outputs and results of its staff members, its associated researchers, or from researchers that have used the resources of the Center. These documents are work in progress, and thus, are subject to changes. Their authors welcome feedback and comments of readers.

El contenido de los documentos de la Serie no representa la posición oficial de CERAC, los miembros de sus organismos de dirección o las entidades que proveen apoyo al centro.

The content of the Working Papers Series does not represent CERAC´s points of view, the members of their direction organisms or the entities that provide support to the Center.

Mauricio Sadinle es estudiante de la Maestría en Estadística de la Universidad Nacional de Colombia. Su formación incluye estudios en estadística, demografía, probabilidad, matemáticas y economía. En su ejercicio disciplinar ha desarrollado varios trabajos teóricos en el área de probabilidad y estadística, ha colaborado con la implementación de metodologías estadísticas en el software libre R y ha trabajado en temas de desplazamiento forzado interno. En CERAC trabaja en el estudio de poblaciones en cuanto a la estimación de sus tamaños y su caracterización.

Mauricio Sadinle is student of the Master of Statistics at Universidad Nacional de Colombia. His training includes studies in statistics, demography, probability, mathematics and economy. In his disciplinary exercise he has developed theoretical works in the area of probability and statistics, he has collaborated with the implementation of statistical methodologies in the free software R and he has worked in topics of internal forced displacement. In CERAC he works in the study of populations, the estimation of their sizes and their characterization.

ISSN: 1909-1397
CERAC - Centro de Recursos para el Análisis de Conflictos
Calle 55 núm. 5 -22
Teléfonos: +(571) 217 0200
Fax: +(571) 217 0200 ext. 210
www.cerac.org.co
Bogotá, Colombia

On the Performance of Dual System Estimators of Population Size: A Simulation Study

Mauricio Sadinle*

Conflict Analysis Resource Center – CERAC and Departamento de Estadística, Universidad Nacional de Colombia

December 2008

Abstract

A simulation study is carried out in order to compare the performance of the Lincoln–Petersen and Chapman estimators for single capture–recapture or dual system estimation when the sizes of the samples or record systems are not fixed by the researcher. Performance is explored through both bias and variability. Unless both record probabilities and population size are very small, the Chapman estimator performs better than the Lincoln–Petersen estimator. This is due to the lower variability of the Chapman estimator and because it is nearly unbiased for a set of population size and record probabilities wider than the set for which the Lincoln–Petersen estimator is nearly unbiased. Thus, for those kind of studies where the record probability is high for at least one record system, such as census correction studies, it should be preferred the Chapman estimator.

Key words: Bias, Capture–Recapture, Dual System Estimation, Multinomial Distribution, Variability.

1 Introduction

Dual (record) system estimation or single capture–recapture estimation is used to estimate the size of a closed population with two independent samples. This methodology has been widely used in wildlife studies for estimating abundance parameters [e.g. Kekäläinen et al., 2008, Olsson et al., 2006] and has been proposed for census correction [Wolter, 1986]. The United States Census Bureau applied this method by the 1990 Post–Enumeration Survey [see Breiman, 1994, Hogan, 1993, Alho et al., 1993, Mulry and Spencer, 1991] producing estimations by strata attempting the idea presented by Chandra-Sekar and Deming [1949]. These techniques have also been used for estimating population prevalence in epidemiological studies using data obtained from separate sources or record systems [e.g. Seber et al., 2000, Faustini et al., 2000, Abeni et al., 1994].

There are several appropriate approaches depending on the characteristics of the population and the sampling methods for capturing individuals, for example under population heterogeneity or with multiple record systems. This work is just concerned with the dual system estimation method, i.e., estimation of the size of a closed population with only two record systems leading to dual samples. Nevertheless, the reader is referred to Pollock [2000] for a general discussion, to Seber [2001] for relatively new advances from the perspective of capture–recapture methods and to Chao [2001] for a review of models for closed populations. Also the reader interested in multiple samples or multiple record systems is really encouraged to consult Bishop et al. [1975]

*Researcher, CERAC and Graduate Student, Departamento de Estadística, Universidad Nacional de Colombia. E-mail: mauricio.sadinle@cerac.org.co.

where the methodology of estimation for multiple record systems or multiple recapture through log–linear models is presented.

The aim of this work is to compare the performance of the two most referenced estimators for dual system or capture–recapture estimation. Performance refers to the bias and the variability of the estimators, for a widely set of configurations of population size and record probabilities. The estimators to be compared are the Lincoln–Petersen estimator [see Pollock, 2000] and the modification proposed by Chapman [1951], both presented in Section 2 and compared through a simulation study in Section 3.

2 Dual System Estimators

The dual system estimation model allows the estimation of the size of a closed population N when two independent samples of the population are available, the first one of size n_1 and the second one of size n_2. The following assumptions are required: all individuals have the same probability to be caught in each sample, it is possible to identify the m individuals in both samples and the size of the population is constant. Under these it is possible to obtain an adequate estimation of N using dual system or single capture–recapture estimators. The first developed estimator of N for these conditions is due to Lincoln and Petersen [see Pollock, 2000, Le Cren, 1965] given by

$$\widehat{N}_{LP} = \frac{n_1 n_2}{m} \tag{1}$$

and an estimator of its variance is presented by Bishop et al. [1975] as

$$\hat{\sigma}^2(\widehat{N}_{LP}) = \frac{n_1 n_2 (n_1 - m)(n_2 - m)}{m^3} \tag{2}$$

Even though the Lincoln–Petersen estimator is a maximum–likelihood estimator of the population size, it lacks finite moments because it is possible that $m = 0$ if $n_1 + n_2 < N$. Thus, Chapman [1951] modifies (1) and proposes

$$\widehat{N}_C = \frac{(n_1 + 1)(n_2 + 1)}{m + 1} - 1 \tag{3}$$

Taking n_1 and n_2 as fixed parameters, m follows the hypergeometric distribution. From this Wittes [1972] demonstrates that the Chapman estimator is unbiased if $n_1 + n_2 \geq N$, and that when $n_1 + n_2 \leq N$ its bias is given by

$$E(\widehat{N}_C) - N = -\frac{(N - n_1)!(N - n_2)!}{(N - n_1 - n_2 - 1)!N!} \tag{4}$$

Hence, under the latter conditions the Chapman estimator underestimates the size of the population. Also Wittes [1972] presents an estimator of the variance of the Chapman estimator, given by

$$\hat{\sigma}^2(\widehat{N}_C) = \frac{(n_1 + 1)(n_2 + 1)(n_1 - m)(n_2 - m)}{(m + 1)^2(m + 2)} \tag{5}$$

Nevertheless, it is not clear from these results the effects of varying population size and record probabilities on the direction and size of the bias for the Lincoln–Petersen and Chapman estimators under a sampling method unconditional to the values of n_1 and n_2. This scenario is observed in studies where size of the samples can not be determined by the researcher, but they are random variables.

3 A Simulation Study

In this section it is carried out a simulation study in order to explore the performance of the Lincoln–Petersen and Chapman estimators for the size of a population under capture–recapture or dual system estimation. The

performance of these estimators is explored with respect to their biases and their variabilities, for several values of record probabilities and population size.

Under the assumptions of this method, taking p_1 and p_2 as the theoretical probabilities of being registered in the first and second record systems, respectively, unconditional on the values n_1 and n_2, the vector $(m, n_1 - m, n_2 - m, N - n_1 - n_2 + m)$ follows jointly the multinomial distribution [see Bishop et al., 1975]

$$P_\theta(m, n_1 - m, n_2 - m, N - n_1 - n_2 + m) = \frac{N!\ p_{11}^m\ p_{12}^{n_1-m} p_{21}^{n_2-m} p_{22}^{N-n_1-n_2+m}}{m!(n_1-m)!(n_2-m)!(N-n_1-n_2+m)!} \tag{6}$$

These multinomial probabilities depend on a vector of parameters $\theta = (N, p_{11}, p_{12}, p_{21}, p_{22})$, where the probabilities are obtained, from the assumption of independence of the samples, as $p_{11} = p_1 p_2, p_{12} = p_1(1 - p_2), p_{21} = (1 - p_1)p_2, p_{22} = (1 - p_1)(1 - p_2) = 1 - p_{11} - p_{12} - p_{21}$. Note that for the dual system estimation, the parameters of the multinomial distribution are reduced to $\theta = (N, p_1, p_2)$.

Using a Monte Carlo simulation, 2000 random vectors from a multinomial distribution with parameters θ were generated. The multinomial random vectors were generated conditional to $m \neq 0$, which is a desired condition for this kind of estimations. This is clear from the description of the method in wildlife studies, where n_1 individuals are captured, marked and released, and a second later sample of n_2 individuals is obtained where m individuals are recaptured. Intuitively the proportion of recaptures in the second sample should be near the proportion of the first sample in the population, i.e. $m/n_2 \approx n_1/N$, and the Lincoln–Petersen estimator is obtained as $\widehat{N}_{LP} = n_1 n_2/m$. Thus, it is clear that the basis for this method is the count m of recaptured individuals.

For generating random vectors from the multinomial distribution (6) above, it is needed to fix three parameters: N, p_1 and p_2. The mean, the quantiles 2.5% and 97.5% of the simulated distribution of the Lincoln–Petersen and Chapman estimators are shown in Figure 1 and Figure 2, respectively, as a function of p_1 for fixed values $N = \{20, 100, 500\}$, $p_2 = \{0.1, 0.5, 0.9\}$ and for 300 uniformly distributed values of p_1 between 0.01 and 0.99. Note that the roles of p_1 and p_2 are symmetric. The performance of those estimators is reported relative to the true population size, i.e. it is reported the performance of $\widehat{N}/N$ for each estimator. This is made in order to compare the relative magnitude of the bias and the variances through different population sizes. Figure 3 and Figure 4 show the mean, the quantiles 2.5% and 97.5% of the simulated distribution of $\widehat{N}/N$, for the Lincoln–Petersen and Chapman estimators respectively, letting N vary into the set of values $\{20, 21, \ldots, 1000\}$ fixing $p_1 = \{0.1, 0.3, 0.7\}$ and $p_2 = \{0.1, 0.5, 0.9\}$. In all those graphs, as the mean of the distribution (solid line) is closer to 1, the bias is lower. Also, notice that the smaller is the area between the 2.5% and 97.5% percentiles of the distribution (dotted lines), the lower is the variability of the estimator. This simulation study was carried out using the statistical software R [R Development Core Team, 2008].

Figure 1 shows that the Lincoln–Petersen estimator has a poor performance with small probabilities and small population size. For some configurations of the parameters N and p_2 it also presents a poor bias performance for medium values of p_1. As a general overview, the Lincoln–Petersen estimator seems to be nearly unbiased for a compromise between N, p_1 and p_2 under the multinomial distribution of $(m, n_1 - m, n_2 - m, N - n_1 - n_2 + m)$. For example, as it can be seen in Figure 1, for $N = 20$, $p_2 = 0.1$ this estimator underestimates the true value N except for large values of p_1, i.e. $p_1 > 0.7$, where it becomes nearly unbiased. Also, for $N = 100$, $p_2 = 0.1$ and $N = 20$, $p_2 = 0.5$ it underestimates for values of p_1 near 0, but for values $p_1 > 0.15$ it overestimates until p_1 around 0.8, where it becomes nearly unbiased. Even though this estimator is nearly unbiased for $N = 500$, for almost all values of p_1 and p_2, it presents bias for very small values of p_1, but its bias decreases as p_1 increases. As it is shown in Figure 1, this estimator has a similar performance in terms of its bias for $N = 500$, $p_2 = 0.1$ and $N = 100$, $p_2 = 0.5$ due it presents a negative bias for very small values of p_1, later it presents positive bias, and finally it becomes nearly unbiased. For $N = 500$, $p_2 = 0.5$; $N = 100$, $p_2 = 0.9$ and $N = 500$, $p_2 = 0.9$ this estimator is nearly unbiased for almost all values of p_1, except when it is very near to 0, i.e. $p_1 < 0.1$. Thus the performance of the bias of this estimator can be compensated by increasing the record probabilities when there are small populations, as a compromise between N, p_1 and p_2.

In Figure 2 it can be seen that the bias of the Chapman estimator also depends of a compromise between

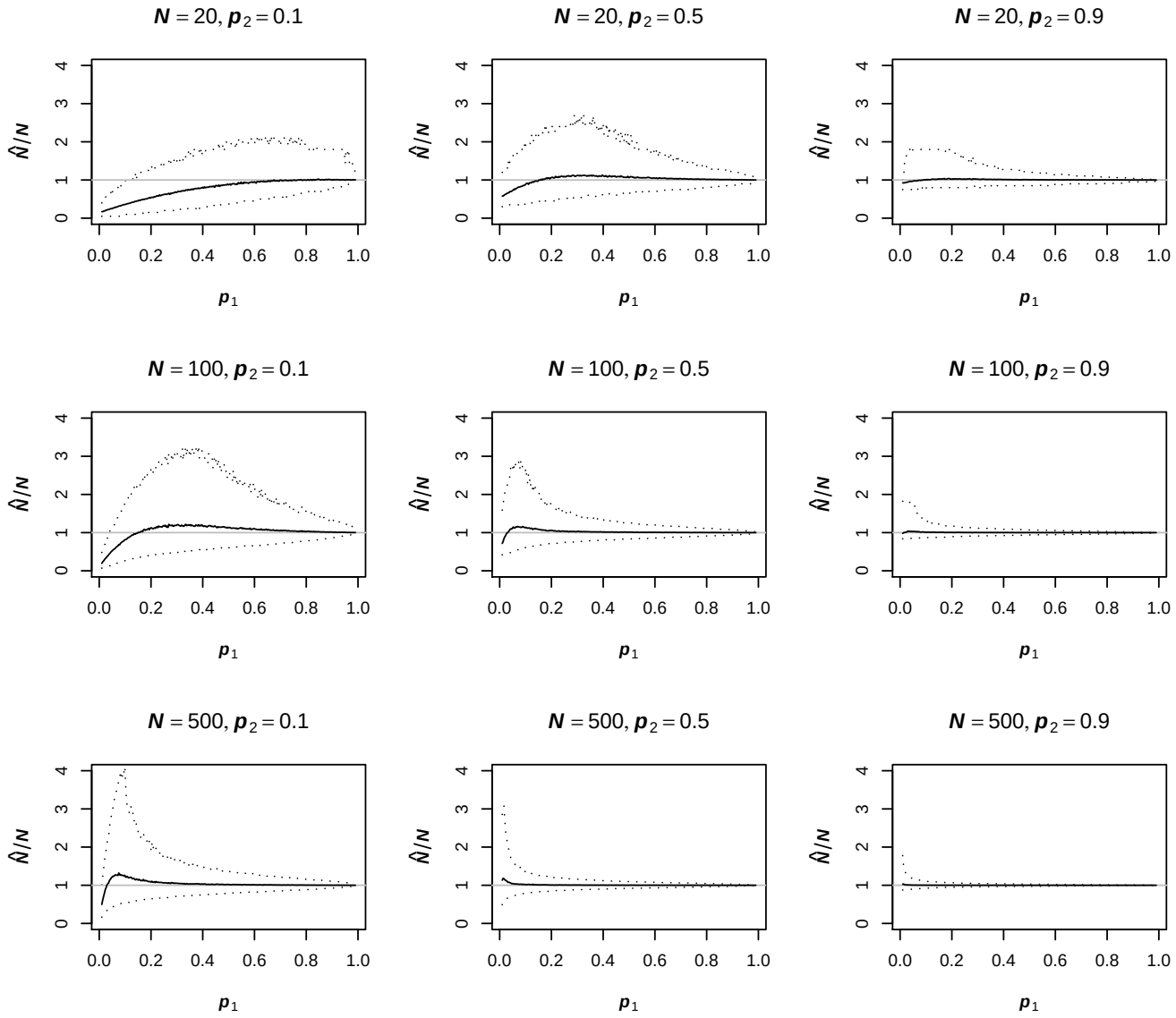


Figure 1: Relative performance of the Lincoln–Petersen estimator as a function of p_1 for fixed N and p_2. The solid line, the upper and lower dotted lines denote the mean, the 97.5% and 2.5% quantiles of the simulated distribution of $\widehat{N}_{LP}/N$.

N, p_1 and p_2. However, the bias of this estimator is never positive, as it is known since the work of Wittes [1972] for a sampling method conditional to n_1 and n_2 which leads to the hypergeometric distribution of m. In comparison with the Lincoln–Petersen estimator, the Chapman estimator is nearly unbiased, for the same N and p_2, since smaller values of p_1, except when $N = 20$ and $p_2 = 0.1$ where it is unbiased only for $p_1 \approx 1$. For example, this estimator is nearly unbiased for $N = 100$, $p_2 = 0.1$ and $N = 20$, $p_2 = 0.5$ when $p_1 > 0.4$, for $N = 500$, $p_2 = 0.1$ and $N = 100$, $p_1 = 0.5$ when $p_1 > 0.1$, for $N = 20$, $p_2 = 0.9$ when $p_1 > 0.2$, and for $N = 500$, $p_2 = 0.5$; $N = 100$, $p_2 = 0.9$ and $N = 500$, $p_2 = 0.9$ the Chapman estimator is unbiased for almost all p_1, except for $p_1 < 0.05$, i.e., the Chapman estimator is nearly unbiased for a range of values of p_1 wider than the range of values for which the Lincoln–Petersen estimator is nearly unbiased.

Comparing Figure 1 and Figure 2 shows that the variance of the distribution of the Lincoln–Petersen estimator is larger than the variance of the distribution of the Chapman estimator. Nevertheless, for both estimators, their variances do not necessarily decreases as p_1 increases, but under conditions of nearly unbiasedness this holds true.

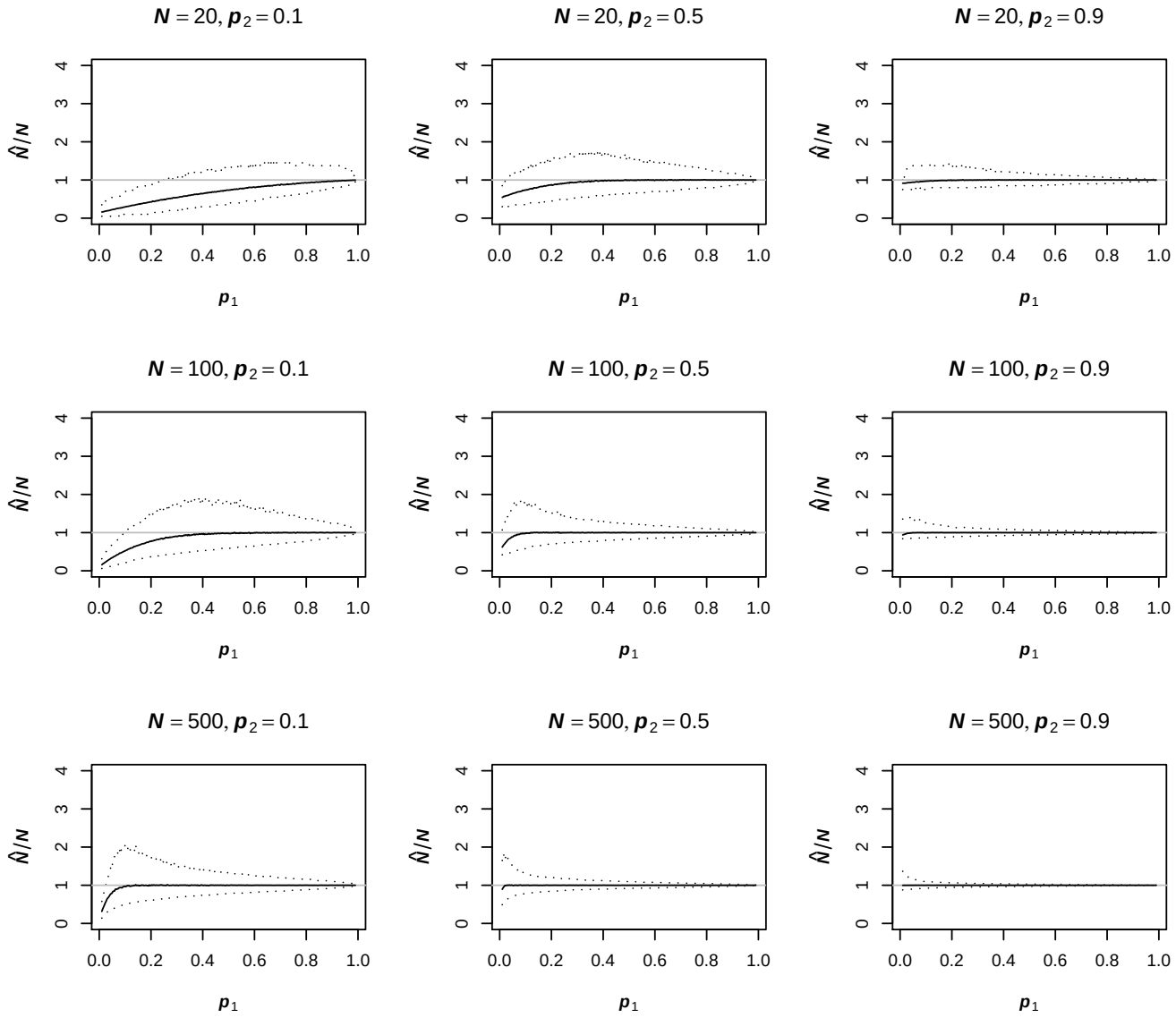


Figure 2: Relative performance of the Chapman estimator as a function of p_1 for fixed N and p_2. The solid line, the upper and lower dotted lines denote the mean, the 97.5% and 2.5% quantiles of the simulated distribution of $\widehat{N}_C/N$.

The performance exposed above is also shown in Figure 3 and Figure 4. For example, for all the values of N reported with $p_1 = p_2 = 0.1$, the variability of $\widehat{N}_{LP}$ is greater than the variability of $\widehat{N}_C$, and for the Lincoln–Petersen estimator the bias goes from negative to positive and it approaches zero very slowly as N increases, in contrast with the Chapman estimator, which is nearly unbiased for $N > 500$ with the same record probabilities. In general, for small record probabilities, the performance of the Chapman estimator is better in bias and variance. This also can be shown by the comparison of Figure 3 and Figure 4 for $p_1 = 0.3, p_2 = 0.1$. Also for the other configurations of p_1 and p_2 reported, the variances of the Chapman estimator are lower than the variances of the Lincoln–Petersen estimator. However, the performance of both estimators is very similar for high record probabilities, for all values of N reported, as it can be shown for $p_1 = 0.7, p_2 = 0.9$.

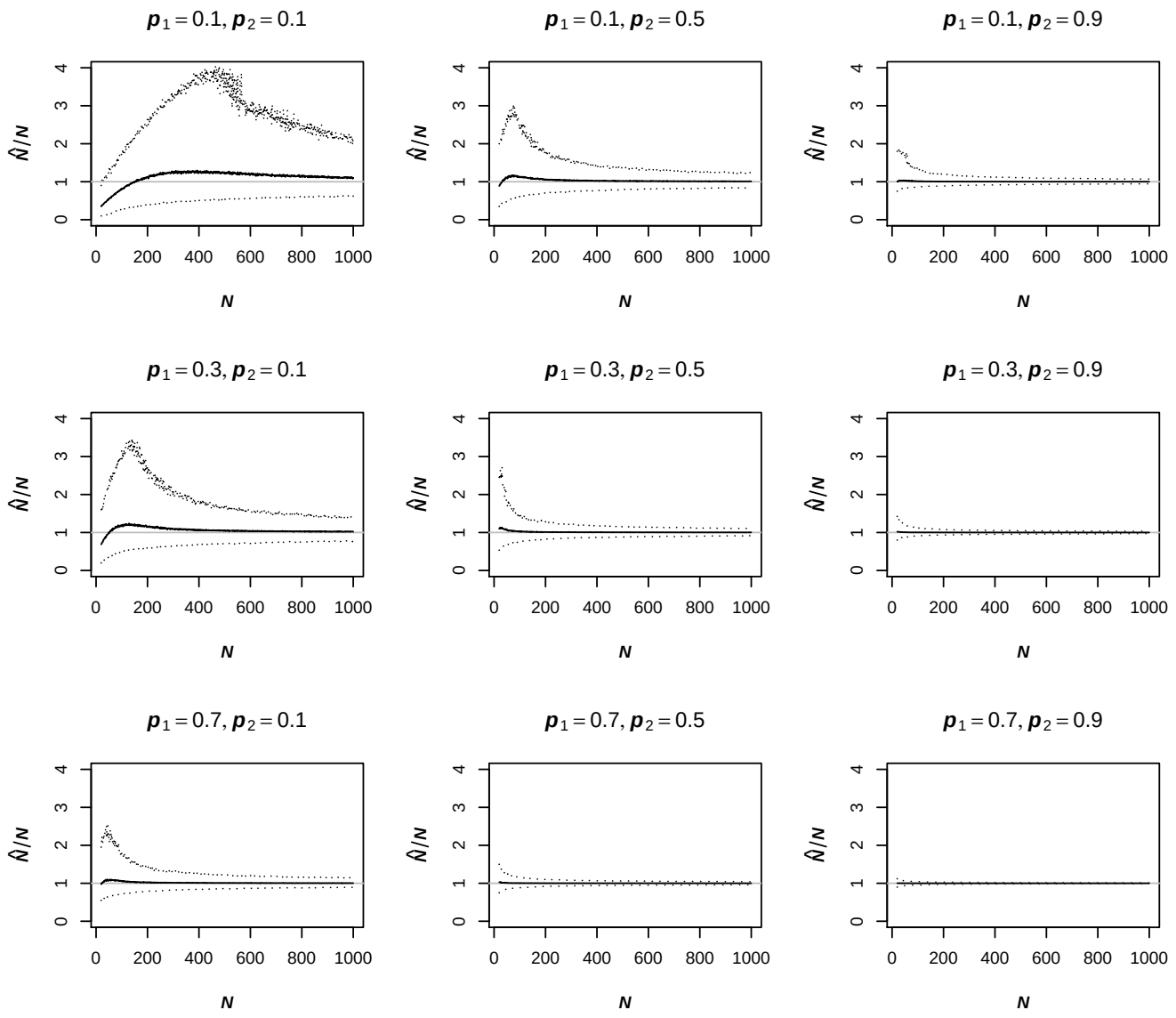


Figure 3: Relative performance of the Lincoln–Petersen estimator varying the size of the population for fixed p_1 and p_2. The solid line, the upper and lower dotted lines denote the mean, the 97.5% and 2.5% quantiles of the simulated distribution of $\widehat{N}_{LP}/N$.

4 Conclusion

Unless both record probabilities and population size are very small, the Chapman estimator presents a better performance in comparison with the Lincoln–Petersen estimator. In general, the Chapman estimator has a lower variance and it reaches the unbiasedness for a set of parameters wider than the set of parameters for which the Lincoln–Petersen estimator is unbiased. Also a disadvantage of the Lincoln–Petersen estimator is presented by its positive bias for a large set of configurations of record probabilities and population size, while the Chapman estimator is never positively biased, and thus it can be considered as conservative. Thus, for those kind of studies where the record probability is high for at least one record system, it should be preferred to use the Chapman estimator. Such case is presented in census corrections through Post–Enumeration Surveys, where the probability to be registered in the census is high.

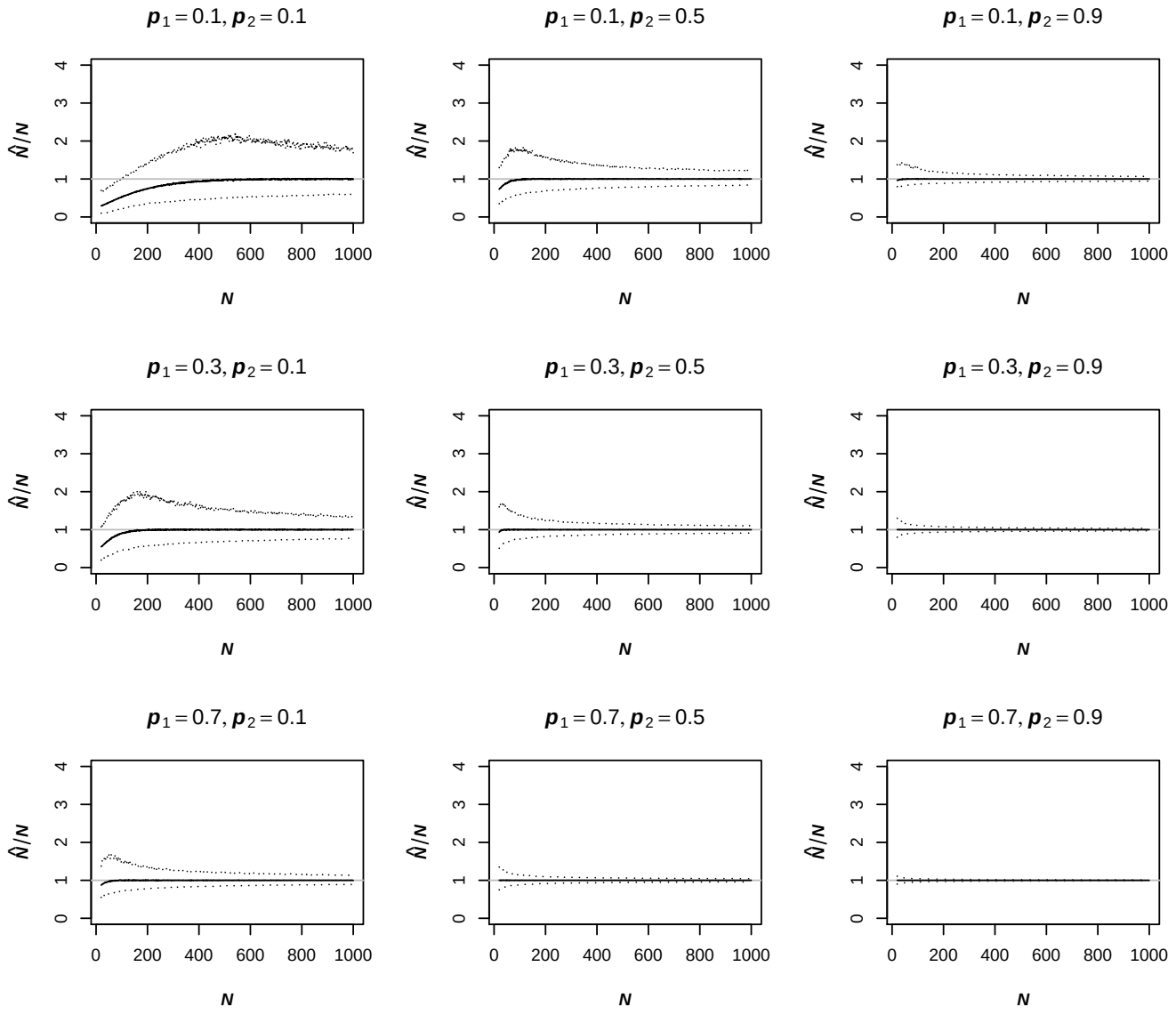


Figure 4: Relative performance of the Chapman estimator varying the size of the population for fixed p_1 and p_2. The solid line, the upper and lower dotted lines denote the mean, the 97.5% and 2.5% quantiles of the simulated distribution of $\widehat{N}_C/N$.

5 Acknowledgments

The author thanks to Jorge A. Restrepo for inspiring this research and his helpful comments.

References

Damiano D. Abeni, Giovanna Brancato, and Carlo A. Perucci. Capture-Recapture to Estimate the Size of the Population with Human Immunodeficiency Virus Type 1 Infection. *Epidemiology*, 5(4):410–414, July 1994.

Juha M. Alho, Mary H. Mulry, Kent Wurdeman, and Jay Kim. Estimating Heterogeneity in the Probabilities of Enumeration for Dual-System Estimation. *Journal of the American Statistical Association*, 88(423):1130–1136, September 1993.

Yvonne M. Bishop, Stephen E. Fienberg, and Paul W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. The MIT Press, 1975.

Leo Breiman. The 1991 Census Adjustment: Undercount or Bad Data? *Statistical Science*, 9(4):458–475, November 1994.

C. Chandra-Sekar and W. Edwards Deming. On a Method of Estimating Birth and Death Rates and the Extent of Registration. *Journal of the American Statistical Association*, 44(245):101–115, March 1949.

Anne Chao. An Overview of Closed Capture-Recapture Models. *Journal of Agricultural, Biological, and Environmental Statistics*, 6(2):158–175, June 2001.

D. G. Chapman. Some properties of the hypergeometric distribution with applications to zoological simple censuses. *University of California Publications in Statistics*, 1:131–160, 1951.

A. Faustini, V. Fano, M. Sangalli, S. Ferro, L. Celesti, P. Contegiacomo, V. Renzini, and C. A. Perucci. Estimating Incidence of Bacterial Meningitis with Capture-Recapture Method, Lazio Region, Italy. *European Journal of Epidemiology*, 16(9):843–848, 2000.

Howard Hogan. The 1990 Post-Enumeration Survey: Operations and Results. *Journal of the American Statistical Association*, 88(423):1047–1060, September 1993.

J. Kekäläinen, T. Niva, and H. Huuskonen. Pike predation on hatchery-reared Atlantic salmon smolts in a northern Baltic river. *Ecology of Freshwater Fish*, 17:100–109, 2008.

E. D. Le Cren. A Note on the History of Mark-Recapture Population Estimates. *The Journal of Animal Ecology*, 34(2):453–454, June 1965.

Mary H. Mulry and Bruce D. Spencer. Total Error in PES Estimates of Population. *Journal of the American Statistical Association*, 86(416):839–855, December 1991.

A. Olsson, D. Emmett, D. Henson, and E. Fanning. Activity patterns and abundance of microchiropteran bats at a cave roost in south-west Madagascar. *African Journal of Ecology*, 44:401–403, 2006.

Kenneth H. Pollock. Capture-Recapture Models. *Journal of the American Statistical Association*, 95(449): 293–296, March 2000.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

G. A. F. Seber. Some New Directions in Estimating Animal Population Parameters. *Journal of Agricultural, Biological, and Environmental Statistics*, 6:140–151, 2001.

George A. F. Seber, John T. Huakau, and David Simmons. Capture-Recapture, Epidemiology, and List Mismatches: Two Lists. *Biometrics*, 56(4):1227–1232, December 2000.

Janet T. Wittes. On the Bias and Estimated Variance of Chapman's Two-Sample Capture-Recapture Population Estimate. *Biometrics*, 28(2):592–597, June 1972.

Kirk M. Wolter. Some Coverage Error Models for Census Data. *Journal of the American Statistical Association*, 81(394):338–346, June 1986.